

## SYSTEM AND METHOD FOR PREDICTION OF BEHAVIOR OF COMPLEX SYSTEMS

### RELATED APPLICATIONS

5 This application is related to co-pending application Serial No. 09/333,172,  
filed June 14, 1999, which is a continuation-in-part of application Serial No.  
08/073,929, filed June 8, 1993, now issued as Patent No. 5,912,993. The subject  
matter of these applications is incorporated herein by reference.

### COMPUTER APPENDIX

A Computer Appendix containing computer program source code for  
programs described herein has been submitted concurrently with the filing of this  
application. The Computer Appendix will be converted to a Microfiche Appendix  
pursuant to 37 C.F.R. 1.96(b). The Computer Appendix, which is referred to as  
15 "Microfiche Appendix A", is incorporated herein by reference in its entirety. The  
Computer Appendix contains material that is subject to copyright protection. The  
copyright owner has no objection to the facsimile reproduction by anyone of the  
patent document or patent disclosure, as it appears in the Patent and Trademark  
Office patent file or records, but otherwise reserves all copyright rights.

### FIELD OF THE INVENTION

The present invention relates generally to a computer-based system and  
method for organization and prediction of behavior in complex systems. More  
particularly, the present invention relates to a system and method for minimizing the  
25 number of parameters required to model the complex system.

### BACKGROUND OF THE INVENTION

Numerous algebraic modeling methods have been proposed in efforts to  
organize the properties of complex systems in order to control and/or predict their  
behavior. Examples of applications of modeling techniques to complex systems  
include economic modeling of securities, inventories, cash flow, sales, and  
marketing, manufacturing and systems control, and scientific applications to spectral

analysis. In many such systems, while the real number of underlying variables that describe the system properties may be small, these variables are unknown. Because of the complexity of the data and the presence of sampling errors, the model can end up with too many parameters, the quantity of which can equal, or even exceed, the  
5 number of data points. This is a common problem in nonparametric analysis, where using too many parameters leads to large statistical uncertainty and biases in the derived model parameters, and in the correlations among them.

In the area of financial prediction, numerous methods have been proposed that are based upon covariance matrix models. For example, in Patent No. 5,444,819, of Negishi, an economic phenomenon predicting and analyzing system using a neural network is described. Learning data is input into the network, including past trends, patterns of variations, and the objective economic phenomenon corresponding to the past data. The hidden layer acts as a number of covariance matrices, categorizing the data in an attempt to identify a small number  
15 of principal variants or components, ideally reducing the number of variants to be considered in the prediction of moving averages.

This neural network of Negishi is an attempt at applying the principle of "minimum complexity", also called "algorithmic information content." This principle is a manifestation of Ockham's razor -- to minimize the number of parameter required to fit the system. Minimum complexity enables an efficient representation of the complex system and is the best way to separate a signal from noise. (For purposes of this application "signal" means the desired information, which can be financial data or other information to be extracted from an input containing an excess of information, much of which can be considered superfluous  
25 background noise.) If the signal can be adequately represented by a minimum of P parameters, addition of another parameter only serves to introduce artifacts by fitting the noise. Conversely, the removal of too many parameters can result in an improper representation of the system, since adequate fitting of a model to the system requires a minimum of P parameters.

~~While minimum complexity has a clear theoretical advantage, it can be computationally intensive, making it difficult to reach a conclusion in a period of~~



time that would permit practical application, unless the parameterization of the system is known in advance. A representation of a system requires a model language that decomposes it into smaller units, and one must choose between a vast number of languages, i.e., means for expressing the algorithm. Even after a language is chosen, the set of all possible parameterizations with that language can become too large to search practically. For example, consider modeling the covariance matrix of a system of  $N$  variables with  $P$  parameters, such as discussed above relative to the Negishi patent. Standard estimates assume that all the elements of the covariance matrix are significant, i.e., each variable is correlated with every other variable. This gives  $N(N+1)/2$  independent elements of the covariance matrix (after accounting for the symmetry of the covariance matrix). A minimum complexity model seek to represent these  $N(N+1)/2$  numbers by a much smaller number  $P$  of parameters. One simple approach would be to zero all but  $P$  of the  $N(N+1)/2$  elements. But the choice of  $P$  elements among  $N(N+1)/2$  is a combinatorially large problem and an exhaustive evaluation of all of the possibilities is not practical. In addition, the covariance matrix must be positive definite, and this constraint further restricts the possible parameterizations. It is clear therefore, that a practical method is needed, which will find a minimum-complexity model without having exhaustively to search all possible parameterizations.

Others in the field have proposed prediction and risk assessment techniques based using covariance matrices, with some developing relatively complex models with a large number of variables, thus producing a computationally-intensive model. See, e.g., Tang, "The Intertemporal Stability of the Covariance and Correlation Matrices of Hong Kong Stock Returns", *Applied Financial Economics*, 8:4:359-65; Nawrocki, "Portfolio Analysis with a Large Universe of Assets", *Applied Economics*, 28:9:1191-98. Others have proposed models in which the number of parameters is so small that one must be concerned about the accuracy of the representation of the system. See, e.g., Hilliard and Jordan, "Measuring Risk in Fixed Payment Securities: An Empirical Test of the Structured Full Rank Covariance Matrix", *Journal of Financial and Quantitative Analysis*, Sept. 1991.

In related application Serial No. 333,172, the inventors disclose a signal and

image reconstruction method which utilizes the minimum complexity principle, in which the method adapts itself to the distribution of information content in the image or signal. Since a minimum complexity model more critically fits the image to the data, the parameters of the image are more accurately determined since a larger  
5 fraction of the data is used to determine each one. For the same reason, a minimum complexity model does not show signal-correlated residuals, and hence provides unbiased source strength measurements to a precision limited only by the theoretical limits set by the noise statistics of the data. In addition, since the image is constructed from a minimum complexity model, spurious (i.e., artificial or numerically created) sources are eliminated. This is because a minimum complexity model only has sufficient parameters to describe the structures that are required by the data and has none left over with which to create false sources. These fundamental parameters are known as Pixon™ elements, which are also described in related Patent No. 5,912,993. Finally, because the method builds a critical model  
15 and eliminates background noise, it can achieve greater spatial resolution than competing methods and detect fainter signal sources that would otherwise be hidden by background noise.

It would be desirable to apply the methods of minimum complexity to a method for prediction of behavior of complex systems where the behavior can be modeled algebraically, and where the computation time is appropriate for practical applications.

#### SUMMARY OF THE INVENTION

It is an advantage of the present invention to provide a method for  
25 minimizing the complexity of algebraic models for modeling the behavior of complex systems.

It is another advantage of the present invention to provide a method for minimizing the complexity of algebraic models used for modeling complex systems in a practical amount of time using a conventional computer.

Yet another advantage of the invention is to provide an accurate measurement and prediction of the properties of complex systems.

Still another advantage of the present invention is to provide a method for minimizing the number of factors that are required for modeling of behavior in a complex system.

5 It is another advantage of the present invention to provide a method for minimizing the number of factors needed to characterize a complex system in a practical amount of time using a conventional computer.

It is yet another advantage of the present invention to provide a method for assessment of risk and volatility in securities and other financial investments.

A system and method are provided for building efficient algebraic models of complex systems by seeking the algebraic model with the minimum complexity, i.e., number of parameters, that adequately describes the properties of the system. The invention identifies an Algebron™ element, which is a fundamental and indivisible unit of information contained in the data. The actual Algebron™ elements that are selected during an iterative process represent the smallest number of such units  
15 required to fit the data.

The Algebron™ system and method utilize a software program stored in a personal computer (PC) to determine the minimum number of factors required to account for the input data by seeking an approximate minimum complexity model] that is achievable in limited period of time using a reasonable number of computational steps. In an exemplary embodiment for estimating covariance in the daily returns of financial securities, the method generates a positive-definite estimate of the elements of a covariance matrix consistent with the input data. However, the method minimizes complexity of the covariance matrix by assuming that the number of independent parameters is likely to be much smaller than the number of elements in the covariance matrix. The Algebron™ method minimizes the number of independent parameters by describing each variable as a linear combination of independent factors and a part that fluctuates independently. The simplest model for the covariance matrix is selected so that it fits the data to within a specified quality as determined by the selected goodness-of-fit (GOF) criterion. In this case the GOF criterion is the logarithm of the likelihood function.

The Algebron™ element is a fundamental and indivisible unit of

information, i.e., the theoretical limit of the information content of the data. The ultimate characterization of the covariance matrix of the object of interest is obtained by extracting all of its Algebron™ elements. While based on a principle much like the Pixon™ method first disclosed in Patent No. 5,912,993, the

5 distinction between Algebron™ elements and Pixon™ elements is that Pixon™ elements are spatially based within an image or signal, and their location plays a role in the description of the model. Algebron™ elements, on the other hand, are single algebraic elements in the model developed for the data that have no spatial position.

### BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be better understood from the following detailed description of some preferred embodiments of the invention, taken in conjunction with the accompanying drawings, in which:

15 Figures 1a and 1b show fits to measured data, where Figure 1a is a plot of a  $\chi^2$ -fit to a twelfth order polynomial signal using prior art methods; and Figure 1b is a plot of an Algebron™ fit according to the present invention; and

Figure 2 is a flow chart of the covariance estimator according to the present invention.

### DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

A computer, such as a fast desktop personal computer ("PC"), is used to generate an efficient, minimum complexity algebraic model for the complex system of interest. In the illustrative example described, the Algebron™ method is used to

25 estimate the covariance of non-uniformly sampled financial return data. The example provided herein describes application of the Algebron™ method to securities for purposes of risk management and forecasting. However, the method is not restricted to securities applications, and may be used for a wide range of evaluations of complex systems which typically include large numbers of variables, including a number of business-related applications such as economic predictions,

sales, marketing, inventories, scientific applications, such as spectral analysis, and industrial applications, such as control systems.

Use of the Algebron™ method for efficient algebraic modeling of complex systems is provided by the two examples given below. In the first example noisy data known to be of polynomial form are analyzed by both conventional fitting methods and by the Algebron™ method. In the second example, the Algebron™ method is used to model the covariance matrix of daily returns of a family of bonds and compared to standard techniques for covariance matrix estimation.

### Example 1:

This example illustrates the use of the Algebron™ method to model noisy data that are known to contain an underlying twelfth order polynomial signal. Figures 2a and 2b show fits to the measured data. In the specific example chosen here, the true signal is the polynomial  $y(x) = 30x^4 - 20x^7 - 5x^8$ , and the noise is normally distributed (Gaussian) with a standard deviation of 0.25. Figure 2a shows the result of a  $\chi^2$ -fit of a twelfth order polynomial to the data using standard methods. From visual inspection, the quality of the fit seems quite good. However, the values determined for the coefficients of  $x^4$ ,  $x^7$ , and  $x^8$ , of -27836.394, 728246.77, and -890971.70 respectively, are very far from the correct value of 30, -20, and -5. In fact, equally large magnitude values for the coefficients of the other absent polynomial powers are also determined. These large and disturbing errors are due to fitting the data with too complex a model. These problems can only be alleviated by imposing a minimum complexity constraint on the model such as is done in the Algebron™ method.

The Algebron™ fit is shown in Figure 2b. Visual inspection of this fit also confirms that it is excellent. In contrast to the standard  $\chi^2$ -fit, however, the Algebron™ fit achieves very good accuracy. The values determined for the coefficients of  $x^4$ ,  $x^7$ , and  $x^8$  are 29.8, -20.4, -4.38, very close to the true values of 30, -20, and -5. In addition, the Algebron™ fit finds no powers of  $x$  that are not present in the signal.

Example 2:

In the second example, the Algebron™ method is used to estimate volatility (covariance matrix) of a family of 132 securities reported over a period of 820 days. With complete data, the covariances of the securities are directly calculable from the returns  $X_\alpha$  and  $X_\beta$  of the individual securities  $\alpha$  and  $\beta$ . To reduce the number of independent elements of the covariance matrix, an analysis procedure is used in which each variable  $X_\alpha$  is described as a linear combination of unknown factors  $f_\beta$ , and a part that fluctuates independently,  $N_\alpha$ , which corresponds to the remaining "noise" associated with the securities:

$$X_\alpha = \sum_{\beta=1}^k \Lambda_{\alpha,\beta} f_\beta + N_\alpha, \quad (1)$$

The  $k$  factors,  $f_\beta$ , are independent of each other, and independent of the noise terms  $N_\alpha$ . The goal of the factor analysis is to determine the minimum set of factors necessary to describe the observations. When this is accomplished, the covariance matrix  $V_{\alpha,\beta}$  is completely described in terms of the loading matrix  $\Lambda_{\alpha,\beta}$ , and additional, independent dispersions  $\sigma_\alpha$ :

$$V_{\alpha,\beta} = \sum_k \Lambda_{\alpha k} \Lambda_{\beta k} + \sigma_\alpha^2, \quad (2)$$

where  $\sigma_\alpha^2$  is the variance of  $N_\alpha$ . Thus, the complexity of the model is minimized by minimizing the number of factors needed to account for the data as well as minimizing the number of nonzero loading matrix coefficients  $\Lambda_{\alpha k}$ .

The data set chosen for this example presents tremendous problems for standard analysis techniques. For example, in the time series of the returns, two-thirds of all of the possible data are missing because of gaps in the trading or the reporting of the returns. Hence direct calculation of the covariance matrix is impossible. Some form for sophisticated algebraic modeling is essential in order to estimate the covariance. In this example, the Algebron™ model of the covariance matrix required 8 factors and a total of 152 off-diagonal, nonzero loading matrix coefficients. This compares to standard factor analysis methods that would use 924



off-diagonal, nonzero loading matrix elements for this number of factors. As with Example 1, the presence of these additional large number of unnecessary model parameters grossly affects the determination of the remaining required parameters. Hence the ten-fold reduction in model complexity afforded by the Algebron<sup>TM</sup> method leads to a tremendous improvement in accuracy.

Figure 2 provides a top level flowchart 100 for covariance estimation using the Algebron<sup>TM</sup> method, with corresponding references to several subroutines of the computer program provided in accompanying Computer Appendix A. In step 2, the input data has been provided and the initial starting condition assumes that there are zero factors  $k$  in the loading matrix  $\Lambda$ , i.e., there is no loading matrix, and the covariance matrix  $V$  consists of diagonal  $S$ . At this point, the sigmas,  $\sigma$ , are set to the sampling standard. The main routine for performing the Algebron<sup>TM</sup> method is provided in the [algebron.f90] subroutine contained in Computer Appendix A. Loop 4-18 begins by the addition of one factor, which in the first iteration means that there is now one factor. In step 6, a line search is performed along the initial direction, which is the  $\Lambda$  direction with the largest negative curvature, i.e., the strongest descent direction in the multivariate saddle point. This step calls up the [linmin.f90] subroutine, which compares the values to the minimization limits. The goal is to minimize the log-likelihood function  $L$  (in the [ml.f90] subroutine) and the covariance matrix  $V$  as follows:

$$L = -2 \ln \Pr(D|M) = \sum_n w_n \left( \ln \|V_n\| + x_n \circ V_n^{-1} \circ x_n \right), \quad (3)$$

where  $n$  sums over the samples, and  $\Pr(D|M)$  is a goodness-of-fit quantity measuring the probability of the data  $D$  given the model  $M$ . The covariance matrix is described in terms of the loading matrix  $\Lambda$ :

$$V_{ij} = \sigma_i \sigma_j \left( \delta_{ij} + \sum_{\alpha} \Lambda_{i\alpha} \Lambda_{j\alpha} \right), \quad (4)$$

where  $\alpha$  sums over the  $k$  factors being considered, and  $\delta_{ij}$  is the Kronecker delta. In step 8, parameters are culled by truncating all elements below a predetermined

value (subroutine [cull.f90]) according to Equation 5:

$$\Lambda_{i\alpha} \rightarrow 0 \text{ for } \Lambda_{i\alpha}^2 \frac{\left(\sum_n w_n\right)^2}{\sum_n w_n^2} < n_\sigma^2, \quad (5)$$

where the sum is only over points for which a sample of variable  $i$  exists. The covariance matrix is re-scaled for the correct  $\chi^2$  before and after the minimization according to the relationship:

$$V \rightarrow \alpha V; \quad \alpha = \frac{\sum_n w_n x_n \cdot V_n \cdot x_n}{\sum_n w_n m_n}. \quad (6)$$

If there is no direction to meet the criteria, the program exits the loop.

In step 10, the log-likelihood function, the selected goodness-of-fit criterion for this example, is minimized by using the conjugate gradient method, after which another step of parameter culling and re-scaling is performed according to the relationships provided in Equations 5 and 6 (step 12). The constrained conjugate gradient is minimized in step 14 by forcing the components of  $\Lambda$  that were set to zero to remain at zero during the calculation. Another rescaling is performed in step 16, and a comparison is made for goodness-of-fit (subroutine [get\_gof.f90]) (step 18). If improvement in the log-likelihood function  $L$  is significant at the current level, the loop repeats, returning to step 4 of another iteration of the steps 4-18 loop. If the improvement in  $L$  is not significant at the prescribed significance level, or there is no increase in the number of parameters  $k$ , the program will back up to the previous solution, which is saved in memory during each iteration, and the estimation process is terminated (step 20). The estimated covariance matrix is converted to data that can be output to an appropriate display. As will be apparent to those of skill in the art, the above-identified subroutines interface with a number of other subroutines that are included in Computer Appendix A, but which are not

specifically referenced or described in the written description. Further, although the bulk of the code provided is written in FORTRAN-90, it will be apparent to those of skill in the art that other programming languages and operating systems can be used. For example, files are provided within the Appendix to generate the object and  
5 binary code on UNIX machines.

The system and method of the present invention enable the rapid estimation of complex systems by building a minimum complexity model, allowing minimization of the number of parameters used to describe the system. The Algebron™ method enables the use of readily available personal computers to obtain the best possible algebraic model with no loss of key information with a minimum number of computational steps, thus providing accurate models in a period of time that is much shorter than computationally-intensive estimation methods of the prior art. The Algebron™ system and method are applicable to any number of complex systems, including financial, scientific or industrial applications, which can be  
15 modeled using algebraic methods.

It will be evident that there are additional embodiments and applications which are not specifically included in the detailed description, but which fall within the scope and spirit of the invention. The specification is not intended to be limiting, and the scope of the invention is to limited only by the appended claims.

WE CLAIM: